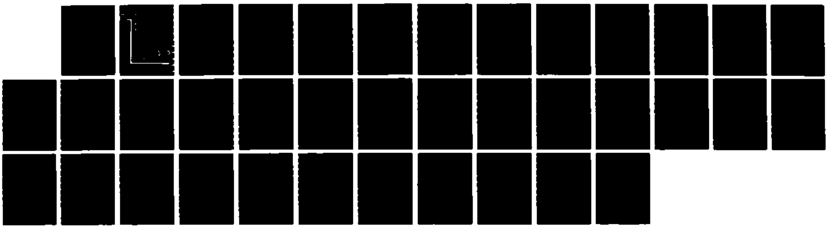
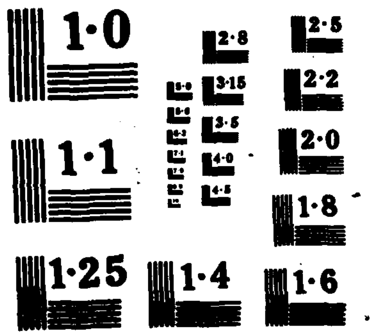
AD-A170 400 PERFORMANCE RATINGS: DESIGNS FOR EVALUATING THEIR VALIDITY AND ACCURACY(U) OLD DOMINION UNIV NORFOLK VA DEPT OF PSYCHOLOGY T L DICKINSON JUL 86 AFHRL-TP-86-15 1/1

UNCLASSIFIED F49620-82-C-0035 F/G 5/9 NL




1·0
2·8
2·5
3·15
2·2
3·5
2·0
1·1
4·0
4·5
1·8
1·25
1·4
1·6

(12)

AFHRL-TP-86-15

AIR FORCE

HUMAN RESOURCES

AD-A170 400

# PERFORMANCE RATINGS: DESIGNS FOR EVALUATING THEIR VALIDITY AND ACCURACY


Terry L. Dickinson

Old Dominion University
Department of Psychology
Norfolk, Virginia 23508


MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601


July 1986

Interim Technical Paper for Period May - September 1983




LABORATORY



AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

## NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

JERRY HEDGE
Contract Monitor

WILLIAM E. ALLEY, Scientific Advisor
Manpower and Personnel Division

RONALD L. KERCHNER, Colonel, USAF
Chief, Manpower and Personnel Division

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

AD-A170400

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>Unclassified | 1b. RESTRICTIVE MARKINGS |
| 2a. SECURITY CLASSIFICATION AUTHORITY | 3. DISTRIBUTION/AVAILABILITY OF REPORT<br>Approved for public release; distribution is unlimited. |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S) | 5. MONITORING ORGANIZATION REPORT NUMBER(S)<br>AFHRL-TP-86-15 |
| 6a. NAME OF PERFORMING ORGANIZATION<br>Old Dominion University — 6b. OFFICE SYMBOL (If applicable) | 7a. NAME OF MONITORING ORGANIZATION<br>Manpower and Personnel Division |
| 6c. ADDRESS (City, State, and ZIP Code)<br>Department of Psychology<br>Norfolk, Virginia 23508 | 7b. ADDRESS (City, State, and ZIP Code)<br>Air Force Human Resources Laboratory<br>Brooks Air Force Base, Texas 78235-5601 |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Office of Scientific Research — 8b. OFFICE SYMBOL (If applicable) AFOSR | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER<br>F49620-82-C-0035 |
| 8c. ADDRESS (City, State, and ZIP Code)<br>Bolling Air Force Base<br>Washington DC 20332 | 10. SOURCE OF FUNDING NUMBERS<br>PROGRAM ELEMENT NO. 62703F; PROJECT NO. 7734; TASK NO. 08; WORK UNIT ACCESSION NO. 22 |

11. TITLE (Include Security Classification)
Performance Ratings: Designs for Evaluating Their Validity and Accuracy

12. PERSONAL AUTHOR(S)
Dickinson, T.L.

| 13a. TYPE OF REPORT | 13b. TIME COVERED | 14. DATE OF REPORT (Year, Month, Day) | 15. PAGE COUNT |
|---|---|---|---|
| Interim | FROM May 83 TO Sep 83 | July 1986 | 34 |

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

| FIELD | GROUP | SUB-GROUP |
|---|---|---|
| 05 | 08 | |
| 05 | 09 | |

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
accuracy
multitrait-multimethod
performance ratings
rating errors
research design

19. ABSTRACT (Continue on reverse if necessary and identify by block number)
Ratings are an important source of information about job performance. For many jobs, objective measures of performance are not available or are impractical to obtain so that ratings are a distorted source of information. The multitrait-multimethod designs have been used to investigate the distortions in ratings. The purpose of using these research designs is to isolate the factors that distort the ratings and to use this knowledge to improve the quality of performance ratings. The goal of the present research was to develop a design that combined both the multitrait-multimethod and person perception designs. Each design was discussed, and examples were presented to illustrate that design. The combination design was used to isolate the influence of rater, ratee, and context factors on the quality of performance ratings. Future research was recommended to understand and utilize these factors to improve performance ratings.

| | |
|---|---|
| 20. DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>[X] UNCLASSIFIED/UNLIMITED [ ] SAME AS RPT. [ ] DTIC USERS | 21. ABSTRACT SECURITY CLASSIFICATION |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL<br>Nancy A. Perrigo, Chief, STINFO Office | 22b. TELEPHONE (Include Area Code) (512) 536-3877 — 22c. OFFICE SYMBOL AFHRL/TSR |

DD FORM 1473, 84 MAR

83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

AFHRL Technical Paper 86-15 July 1986

# PERFORMANCE RATINGS: DESIGNS FOR EVALUATING THEIR VALIDITY AND ACCURACY

Terry L. Dickinson

Old Dominion University
Department of Psychology
Norfolk, Virginia 23508

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed by

Rodger D. Ballentine, Lt Col, USAF
Chief, Productivity and Performance Measurement Function

Submitted for publication by

R. Bruce Gould
Chief, Force Utilization Branch
Manpower and Personnel Division

This publication is primarily a working paper. It is published solely to document work performed.

## SUMMARY

Ratings are often the sole source of information about job performance. However, they are not objective measures; ratings can be invalid or contain inaccuracies. Research designs must be used to isolate the factors that distort the ratings, and subsequently, to improve the quality of the ratings. The multitrait-multimethod and person perception designs have been used to isolate such factors. The goal of the present research was to develop a design that combined both the multitrait-multimethod and person perception designs. Examples were presented to illustrate the combination design, and it was used to isolate the influence of rater, ratee, and context factors on the validity and accuracy of performance ratings. It was recommended that the combination design be used in future research to improve performance ratings.

# PREFACE

This research was conducted under the USAF - SCEEE Summer Faculty Program sponsored by the Air Force Office of Scientific Research. The work was accomplished by the author while working in the Manpower and Personnel Division, Air Force Human Resources Laboratory. It complements the efforts of the Productivity and Performance Measurement Function which is working on a long-term job performance criterion development project.

TABLE OF CONTENTS

LIST OF TABLES

PERFORMANCE RATINGS: DESIGNS FOR EVALUATING THEIR VALIDITY AND ACCURACY

# I. INTRODUCTION

Performance ratings are an important method for measuring and defining human attributes. They have been used in research and applied contexts to describe a diversity of human attributes such as group leadership skills, problem-solving ability and interpersonal skills. In some contexts, performance ratings serve as substitutes for more objective but expensive methods such as work sample testing, whereas in other contexts, ratings are the only practical measures of attributes.

Despite the utility of performance ratings, they must be interpreted with caution. Since they require human judgments, performance ratings are fallible measures. Several distortions in ratings have been identified that illustrate their fallibility, including leniency, halo, and similarity errors (Landy & Farr, 1980). Such distortions limit inferences about human attributes and the amounts of those attributes possessed by the individuals who are rated.

The limitations to inferences have been addressed with research into the validity and accuracy of ratings (DeCotiis & Petit, 1978; Saal, Downey, & Lahey, 1980). The validity of ratings has been investigated with a multitrait-multimethod design (Boruch, Larkin, Wolins, & MacKinney, 1970). The purpose in using such a design was to evaluate performance ratings against criteria that are logical requirements for measures of human attributes. In particular, variance components and intraclass correlations were computed to evaluate the individual differences in performance accounted for by the ratings. The accuracy of ratings has been investigated with a person perception design (Cronbach, 1955). The purpose in using such a design was to compare performance ratings against target ratings that have been specified by the investigator for the research context. In this design, accuracy statistics were computed to describe several discrepancies between the performance and target ratings.

Research on the validity of ratings was stimulated by Lawler's (1967) application of a multitrait-multimethod design. He emphasized that several sources are available for obtaining ratings (e.g., supervisors, peers, and the self) and that these sources may differ in their ratings of performance. Lawler encouraged the application of a multitrait-multimethod approach to compare ratings from several sources. Subsequent research has used a multitrait-multimethod design to investigate formats for

obtaining ratings (Burnaska & Hollmann, 1974), the nature of human attributes (Borman & Dunnette, 1975), and rater training (Borman, 1978).

Research on the accuracy of performance ratings has focused on the effects of rater training (Bernardin & Pence, 1980; Borman, 1977, 1979a; Hedge, 1982; McIntyre, Smith, & Hassett, 1984). Borman (1977, 1979a) introduced the person perception design to assess training to avoid leniency and halo errors. He found that an admonishment to avoid these distortions was successful, but accuracy was not improved. Apparently, raters learned to avoid certain distortions but not how to rate accurately. Other studies have addressed the relationship between the accuracy of ratings and rater attributes such as personality, interests and observational skills (Borman, 1979b; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982).

Although several factors have been investigated as determinants of the validity and accuracy of ratings (cf. DeCotiis & Petit, 1978; Landy & Farr, 1980), no comparison has been made of their influence on both validity and accuracy. The research has compared ratings against criteria for describing individual differences in performance or against target ratings that specify appropriate performance. These factors should be included in a research design that assesses their joint influence on the validity and on the accuracy of ratings. Such a design would employ a multifactor approach to investigate the limits that the factors place on inferences about human attributes.

## II. OBJECTIVES OF THE RESEARCH EFFORT

The goal of this research effort was to develop a design to guide investigations of both the validity and accuracy of ratings. The derived design combined the multitrait-multimethod and person perception designs and utilized the procedures of analysis of variance. Prior to the presentation of the combination design, the multitrait-multimethod and person perception designs will be described to provide a background. Examples will be discussed that illustrate all the designs.

## III. VALIDITY OF PERFORMANCE RATINGS

Performance ratings measure attributes that are assumed to account for performance differences among individuals. Although the attributes are identified and operationally defined through established scientific methodologies such as job analysis and criterion development (McCormick, 1976; Smith, 1976), the assumption should be questioned in most rating contexts. Job analysis and criterion development may produce attributes that

are poorly defined, irrelevant, or redundant with other attributes, and the performance ratings for such attributes will reflect no meaningful differences in individual performance.

Multitrait-multimethod validation is a research strategy for assessing the individual differences accounted for by performance ratings (Kavanagh, MacKinney, & Wolins, 1971). In this strategy, a rating measure is defined as a trait-method unit. A trait is conceived as a human attribute that is conceptually and statistically distinct from other attributes accounting for performance. Some examples of attributes include ability to facilitate group discussions, define acceptable work procedures, and provide negative feedback to others. A method is a procedure for operationally defining traits. Some methods include forced-choice scales, checklist scales, and example-anchored scales. In sum, a rating measure taps a particular trait with a particular methodology.

The trait-method combinations in a research study are determined by the rating context. This context is dictated by the interests of the researcher and the nature of performance. For example, a researcher may use job analysis to define the traits that significantly affect the performance of jet engine mechanics for a commercial airline company, decide to measure that performance with two formats for ratings, and obtain ratings of the mechanics by their immediate supervisor. Thus, the researcher "designs" the multitrait-multimethod investigation.

## Basic Design

Analysis of variance has been used to analyze the ratings from a multitrait-multimethod investigation (Boruch et al., 1970). The basic design includes the three factors of ratees, traits, and methods. As shown in Table 1, the variation in ratings is partitioned into seven sources. The researcher is not concerned with all of the sources of variation in the analyses. The fixed effects of Methods, Traits, and Methods x Traits are usually based on scales of convenience to the investigator and provide no information about validity. For example, two methods may differ because one method employs 5-point scales while the other employs 9-point scales, and two traits may differ because one is more socially desirable. In contrast, the random effects of Ratees, Ratees x Traits, Ratees x Methods, and Error provide information about the validity of the measures. These sources allow inferences about the individual differences among ratees.

The Ratees source of variation indicates the ability of the measures to order the ratees. This ordering can be due to either traits or methods, or both. Of course, the more the measures agree in their ordering of ratees, the more the measures describe individual differences among the ratees. The Ratees source of variation is said to reflect the convergent validity of the measures.

The Ratees x Traits interaction indicates differential ordering of the ratees by the traits. Since the traits should reflect different aspects of performance, the interaction is desirable. In fact, the stronger the interaction, the greater the number of distinct discriminations between the ratees with the traits. The Ratees x Traits source of variation reflects the discriminant validity of the measures.

The Ratees x Methods interaction indicates differential ordering of the ratees with the methods. This differential ordering is undesirable. The methods for rating should not influence the ordering of ratees. Only the traits should determine the ordering of ratees. The Ratees x Methods source of variation reflects the method bias of the measures.

The Error source indicates residual variation due to sampling and measurement errors. The size of this effect relative to the remaining sources of variation suggests the extent of differences between the ratees that cannot be accounted for by the traits and the methods.

The Error mean square may be used to compute F-ratios to establish statistical significance for the remaining sources. However, the F-ratios are based on mean squares with large degrees of freedom, and the critical F-values to establish significance are frequently exceeded when the differences are not practically significant. A more appropriate strategy for assessing the relative variation in ratings explained by the sources is to compare variance components. These components provide a comparison of the relative sizes of convergent validity, discriminant validity, method bias, and error, while controlling for degrees of freedom. For a single research study, the variance components may be compared directly. However, since the variance component due to Error would differ from study to study, comparisons of the variance components from several studies is not appropriate. Rather, ratios of the variance components are formed to generate intraclass correlation coefficients. These ratios are expressed as a source's component divided by the sum of all variance components. Each ratio reflects the proportion of variance accounted for by that source

relative to the variation accounted for by all sources.

## Computations

The computations associated with a multitrait-multimethod design may be accomplished in several ways. First, the computations may be conducted directly on the ratings that are obtained in the investigation. These computations use the sum of squares formulas that are traditionally employed in the analysis of variance (e.g., Kirk, 1968, pp. 239-240).

Another computational strategy involves the use of the variance-covariance matrix among the measures (Stanley, 1961). This matrix can be used to compute the sum of squares for the various random effects of interest in the multitrait-multimethod investigation. This computational strategy has the advantage of displaying the contributions of each of the measures to the ordering of the ratees. It is directly related to the use of the correlational matrix among the measures in a multitrait-multimethod investigation (Campbell & Fiske, 1959; Kavanagh, MacKinney, & Wolins, 1971).

## Example

An issue of research in performance measurement is the choice of a method for obtaining performance ratings (Schwab, Heneman, & DeCotiis, 1975). All methods are not equally desirable. Methods should be compared in terms of the individual differences that each accounts for in the ratings. One method is preferred to others if the ratings that are obtained with that method display more discriminant validity and less bias in ordering the ratees.

As an illustration, suppose an investigator needs rating scales for research on the performance of test administrators. The investigator has defined three traits and collected the items for constructing the rating scales (e.g., Dickinson & Zellinger, 1980). However, the investigator still needs to specify a method for obtaining the ratings. The decision regarding a method has been narrowed to a choice between example-anchored scales (Taylor, 1968) and checklist scales (Landy & Trumbo, 1980). To aid in decision making, the investigator has collected data in a multitrait-multimethod design. The data are displayed in Table 2. The analysis of the data that were used by the investigator in making this choice is presented below.

The data were collected from a group of raters who viewed videotapes of 10 test administrators who were played by actors according to 10 scripts of performance. The tests that were given by the administrators were the same in each videotape. However, the performance of the administrators on the traits varied across the videotapes. The group of raters viewed each tape, discussed the performance of that administrator, and rated performance on each of three traits using the example-anchored and checklist methods for rating.

The investigator employed a traditional formulation in conducting an analysis of the variation in the ratings. A summary of the analysis is shown in Table 3. The variance components and intraclass correlation coefficients indicate that the measures can be used to order ratees with substantial validity and with little bias due to the method for rating. Convergent validity and discriminant validity account for approximately two-thirds of the variation that determines the ordering of ratee performance. The example-anchored and checklist methods for rating have little influence on the ordering of ratees. Both are equally desirable, and the investigator can choose either of the two methods on the basis of the results. Additional research or practical considerations must guide the choice between the two methods.

## Beyond the Basic Design

The basic design for a multitrait-multimethod investigation can be expanded to research the factors that distort the validity of ratings. Several theoretical models are available to guide such research (DeCotiis & Petit, 1978; Landy & Farr, 1980; Wherry & Bartlett, 1982). The models describe factors ranging from the ability and motivation of raters to organizational policies concerning the use and purpose of the ratings.

To expand the example, suppose the investigator decomposes the decision from a choice between two methods to a choice between two methods that can be used to collect ratings for two purposes. The investigator has collected data from two groups with the basic multitrait-multimethod design. One group was told that the purpose for the ratings is to research the validity of the tests for selecting new employees, while the second group was told that the purpose is to motivate employees by rewarding or punishing them for their past performance. Finally, the

investigator collected only five ratings from each group. Suppose the videotapes were randomly assigned to the two groups such that the research group viewed videotapes one through five, while the motivation group viewed the remaining videotapes.

A four-factor design was used to analyze the ratings that were collected by the investigator (cf. Winer, 1971, pp. 539-546). The design has factors of Purposes, Ratees nested within Purposes, Traits, and Methods. The psychometric interpretations for the sources of variation are summarized in Table 4.

The expanded multitrait-multimethod design includes both fixed and random effects. As with the basic design, only the random effects in the expanded design provide inferences about individual differences among ratees. For example, the Ratees within Purposes effect represents the ability of the measures to order the ratees in a manner that includes both purposes for the ratings. This effect is a pooling of the Ratee effects available from the two purposes for ratings. This pooled effect includes variation due to convergent validity and the interaction of convergent validity with purpose for the ratings. Unfortunately, the nesting of ratees prevents separating the variation due to convergent validity from its interaction. The decision by the investigator to design the research with ratees nested within purpose groups produces this confounding. Similarly, the variation due to discriminant validity and method bias cannot be separated from their interaction with purpose for the ratings.

A summary of the analysis is shown in Table 5. The expanded research design suggests that purpose for the ratings has little influence on the multitrait-multimethod properties of the ratings. Convergent and discriminant validity again account for substantial differences in the ratings of performance. Little method bias is present; both methods of rating are equally desirable. Purpose for the ratings influences only the raters' use of the scales to describe amounts of the attributes. In particular, the ratees were rated higher on trait number two when the purpose for the ratings was research than when the purpose was for motivation. Since that trait reflects the "warmth" versus "coldness" of the administrator, the investigator suspects that the raters valued this attribute highly in test administrators and believes that the raters may have been emphasizing high standards of rapport with the examinees on the part of the test administrators.

## IV. ACCURACY OF PERFORMANCE RATINGS

Accuracy statistics have been described as the most appropriate criteria for assessing the distortions in ratings (Borman, 1978). Although other statistics are available, most lack a meaningful standard for defining distortions (Saal et al., 1980). In contrast, the person perception design for investigating accuracy requires the development of a standard. The standard is usually a set of target ratings that specifies the performance scores of ratees on several attributes.

Target ratings can be developed from the judgments of experts or other decision-making groups or from objective measures. For example, psychologists have rated the performance of actors as displayed in videotapes. These expert ratings were averaged to define the target ratings (Borman, 1979b). Supervisory ratings of performance have also been used to define target ratings in assessing the accuracy of self-ratings (Mabe & West, 1982). Finally, life history information and paper-and-pencil tests have been used as objective measures to develop target ratings for assessing the accuracy of ratings of interviewee performance (Cline & Richards, 1960).

### Cronbach's Formulation

The overall accuracy of a rater is defined as the sum of squared discrepancies between the rater's performance ratings and the target ratings for the ratees. Cronbach (1955) argued convincingly that overall accuracy should be broken down into four statistics that are mathematically independent components of overall accuracy.

Elevation is the component of accuracy due to the mean of the performance ratings given by a rater for the group of ratees and the set of attributes. The rater whose mean is close to that of the target ratings will tend to rate the performance of the ratees more accurately. Although Cronbach (1955) stated that elevation describes the way a rater uses the rating scale, this statistic is useful for describing the accuracy of the rater in judging the overall performance of a group of ratees (Murphy et al., 1982).

Differential elevation is the component of accuracy

associated with the mean ratings that a rater gives the ratees on the set of attributes. In some rating contexts, these mean ratings for the set could indicate the overall job performance of ratees. This component of accuracy reflects a rater's ability to order ratees in comparison to their overall differences as specified by the means of their target ratings. Murphy et al. (1982) suggest that this component of accuracy is important for administrative decisions. For example, a supervisor is often required to nominate subordinates for training programs or to choose one for promotion.

Stereotype accuracy reflects the accuracy of a rater in using the attributes to describe the group of ratees. The mean ratings on the attributes given by the rater to the group are compared to the mean ratings given to the group by the expert source. This component of accuracy is important in making administrative decisions. For example, a supervisor may need to diagnose relative strengths and weaknesses of a group of subordinates to choose training programs or other developmental activities for them. These decisions require accurate summary evaluations of subordinates on the attributes of performance.

Finally, the most important component of accuracy is *differential accuracy* (Cronbach, 1955). The target ratings for each ratee are compared to the performance ratings given by the rater. Differential accuracy reflects the rater's ability to rate the individual ratee accurately. In an organizational setting, differential accuracy is important for research purposes and for developing employees. Most research projects utilize the performance ratings of individuals, necessitating that each ratee be described with little distortion in the ratings. Employee development requires accurate feedback about an individual's performance, so that changes that are undertaken for improvement are appropriate to the individual.

## Computations

The computations for the accuracy statistics were presented by Cronbach (1955) in his seminal article. These statistics are oriented to the descriptions of the accuracy of each rater, and so, the underlying research design is not emphasized. Indeed, subsequent research studies have utilized the accuracy statistics as measures of the rater's "ability" to perceive others (e.g., Borman, 1979b; Cline & Richards, 1960; Crow & Hammond, 1957). As a consequence, little attention has been given to the basic design underlying person perception investigations and its extension to other areas of research.

## Basic Design

Analysis of variance can be used to partition the rating variance obtained in person perception investigations. The basic design includes the three factors of rating sources, ratees, and traits. Table 6 displays the seven sources of variation in the basic design, and summarizes the psychometric interpretations of these sources.

The sources for ratings are the rater and the experts who provided the target ratings. The variation in ratings accounted for by Rating Sources reflects elevation accuracy. The larger this source of variation, the larger the difference between the overall mean rating of the rater and that of the experts, and the more inaccurate is the rater.

The Ratees effect indicates the ability of the rating sources to describe differences between ratees across the attributes. This effect can be due to the rater, the expert source for the target ratings, or both. Since the investigator will typically select the ratees to differ from one another on the attributes, the Ratees effect should account for substantial variation in the ratings. However, the more the rater agrees with the target ratings, the greater the Ratees effect. The rater who is accurate in ordering the ratees, compared to the expert source, enhances the convergent validity of the ratings.

The fixed effect of Traits reflects the relative amounts of the performance attributes possessed by the group of ratees. The investigator designs this effect into the research with the choice of the rating context and the selection of the ratees. The rating context usually includes attributes that differ in their social desirability and, consequently, some attributes will have greater value to the rater than others. Furthermore, the ratees who are selected by the investigator may be chosen to have unequal amounts of the attributes. If the expert source for ratings provides target ratings that confirm the investigator's intentions, it follows that the Traits effect is likely to account for variation in the ratings.

The Rating Sources x Ratees interaction reflects differential elevation accuracy, and it indicates differential ordering of the ratees by the rater, compared to the expert source for ratings. This differential ordering is undesirable.

An accurate rater should order the ratees in a manner similar to that ordering provided by the expert source. Since the expert source serves as the standard for defining the differences between ratees, the effect can also be considered a reflection of differential convergent validity. A rater may describe more or fewer differences between the ratees in assessing their performance on the set of attributes. The larger the interaction, the more inaccurate is the rater in ordering the ratees.

The Rating Sources x Traits interaction indicates the stereotype accuracy of the rater. An accurate rater should agree with the expert source in the relative amounts of the attributes reflected in the group of ratees. The larger this interaction, the more inaccurate the rater.

The Ratees x Traits interaction reflects the extent of individual differences on the attributes perceived by the rater and the expert source. Since the researcher should select the ratees for the investigation, the differential ordering of the ratees on the attributes can be determined by the researcher. Of course, this assumes that the target ratings are close to the intended performance scores for the ratees (Borman, 1979a). For example, the researcher can construct videotapes of actors who play ratees. Then, the performance of ratees can be acted out in a manner which represents scaled amounts of the attributes. If the investigator selects ratees who differ in their ordering on the traits, then discriminant validity will explain variation in the ratings. Moreover, the more the rater's ratings match those of the expert source, the stronger will be the interaction, and the more accurate will be the rater.

The Rating Sources x Ratees x Traits interaction reflects the differential accuracy of the rater. This is the ability of the rater relative to the expert source to describe individual differences among the ratees. This interaction is undesirable. The rater who is accurate should agree with the expert source on the differences among the ratees. If the rater disagrees with the expert source, the rater will possess more or less discriminant validity than the expert source. Since the target ratings serve as a standard, this differential discriminant validity is undesirable.

## Computations

The sums of squares that are obtained from the analysis of the variance in ratings are closely related to the accuracy statistics developed by Cronbach (1955). The accuracy statistics are contrasts between effects in the analysis of variance design.

Each accuracy statistic is a contrast of effects of the rater to those of the expert source for ratings. Of course, an effect is a linear combination of means, and such combinations are used to compute sums of squares in the design.

## Combination Design

The person perception design for the investigation of accuracy can be combined with the multitrait-multimethod design. The combined design includes the four factors of rating sources, ratees, traits, and methods. In essence, the person perception design has been expanded to include more than one method for obtaining performance ratings, while the multitrait-multimethod design has been expanded to include more than one source for the ratings. As shown in Table 7, the combined design includes the sources and psychometric interpretations of each separate design, as well as several other sources, with their psychometric interpretations.

The Rating Sources x Ratees x Methods interaction reflects the differential ordering of the ratees provided by the rater using the designated methods for rating compared to the ordering provided by the expert source using the same methods. This differential ordering is undesirable. Regardless of the method for rating, an accurate rater should order the ratees in a manner similar to that ordering provided by the expert source. Of course, the expert source for ratings may order ratees differently depending on the method for rating. Since the expert source serves as the standard for defining the differences between ratees, this result can be considered a method bias in the target ratings. However, a logical property for a standard is that it not contain method bias. The target ratings should serve to evaluate the rater's ability to describe ratees, regardless of the method for rating. Hopefully, the investigator can design the research such that the target ratings will be relatively free of method bias.

The Rating Sources x Traits x Methods interaction indicates the accuracy of the rater in using the attributes to describe the group of ratees by the methods. If the investigator designs the research such that the target ratings contain no method bias, this interaction suggests that the rater uses the attributes to describe the performance of the group differently with each method for rating. This interaction is again undesirable. No component of a rater's accuracy should depend on the method for

rating.

Finally, the Ratees x Traits x Methods interaction reflects the influence of the method for the ordering of ratees on the attributes summed over the rater and the expert source. This interaction is also undesirable. The ordering of ratees on the traits should not depend on the method for obtaining ratings. If the investigator designs the research such that the expert source orders the ratees on the traits in the same manner, regardless of the method used, the interaction is determined by the rater's inability to use the methods similarly. This differential use of the methods reflects differential discriminant validity by the rater, and it indicates inaccuracy by the rater. The rater should order the ratees on the attributes regardless of the method that the rater uses for making ratings.

## Example

Consider an extension of the issue of the choice of a method for obtaining performance ratings. Although methods should be compared in terms of the individual differences in ratings that each method accounts for, another aspect is the accuracy with which the rater can use the methods to describe individual differences in the ratings. The multitrait-multimethod design assumes that a method is to be preferred if it influences the ordering of ratees less than other methods. The combination design extends the assumption to consider accuracy. A method is preferred if the rater can use it to obtain greater agreement with the expert source for ratings.

To expand the example, suppose that the investigator developed the scripts and videotapes in a series of workshops with a group of experts. The experts were highly familiar with the performance of test administrators. Scripts were modified and actors changed their performance until the experts were in high agreement in their ordering of the ratees with each method for rating. In sum, the investigator designed the target ratings to contain no method bias.

The investigator employed the combination design to evaluate the accuracy of several raters. The results of the analysis of the ratings that were obtained from one rater are shown in Table 8. The data in Table 2 were used for this analysis. Furthermore, assume that the investigator only collected five ratings from each rating source. Suppose that the expert source and rater each viewed and rated the same set of five videotapes selected randomly from the set of 10.

The results of the research indicate that the rater was fairly accurate. Elevation and differential accuracy accounted for little variation in the ratings; both were not statistically significant. The mean of the performance ratings given by the rater for the group of ratees on the set of attributes compared favorably to the mean provided by the expert source. Importantly, the rater agreed for the most part with the expert source on the differences among the ratees. The Rating Sources x Ratees x Traits interaction was negligible in magnitude, which suggests discriminations by the rater comparable to those by the expert source.

The results do suggest some inaccuracies by the rater. Differential elevation accuracy and stereotype accuracy were both statistically significant. For most ratees, the rater and expert source agreed on individual differences across the set of attributes, regardless of method. However, one of the test administrators was given a much greater mean rating by the expert source. This ratee was the only female actor to play a test administrator, and the investigator suspects that sex may explain the greater mean rating. Perhaps, the rater was prejudiced against female test administrators. The Rating Sources x Traits interaction indicated that the rater did not perceive the traits similarly to the expert source. In particular, trait number two was rated as significantly less prevalent by the rater. This trait reflects the "coldness" versus "warmth" of the test administrator, and the investigator suspects that the rater was insensitive to that attribute of test administration.

The investigator was quite pleased that the method for rating had little influence on the ratings. There was no method bias in ordering the ratees shown by the rater or the expert source. The investigator was successful in eliminating method bias in the target ratings, and utilizing the set of attributes, the rater was able to order accurately the group of ratees. For this rater, at least, the investigator is confident that either method for rating performance can be used to obtain accurate ratings of performance. Nonetheless, the investigator does recognize that the ratings obtained with the example-anchored method cannot be compared in absolute size to those obtained with the checklist method. The Traits x Methods interaction suggests considerable scale bias in measuring the attributes.

## V. DISCUSSION

Several models of the rating process outline variables that influence the accuracy of ratings (DeCotiis & Petit, 1978; Kavanagh, Borman, Hedge, & Gould, 1986; Landy & Farr, 1980). However, none emphasizes the influence of logical requirements for performance measures on accuracy. The research studies that support the models have evaluated accuracy statistics against rater attributes such as personality and training experience. These studies illustrate a myopic research strategy (Cronbach, 1955). They are not connected to meaningful theory about the logical requirements for performance measures.

The combination design can provide a broader strategy for accuracy research. It emphasizes the assessment of accuracy in the framework of logical requirements for performance measures. The investigator can determine conditions under which ratings are obtained including contextual factors, ratees, traits, methods for rating, and sources for target ratings. These conditions allow the investigator to design the amounts of multitrait-multimethod properties into the target ratings. Such logical requirements provide a rich framework for interpreting the accuracy of performance ratings.

Target ratings should be designed to possess the multitrait-multimethod properties found in practice. For example, criterion research consistently shows that job performance is a multidimensional concept (Landy & Trumbo, 1980). There are many routes to success in most work contexts and, so, several attributes are necessary to describe performance. Consequently, the investigator must design the target ratings to possess discriminant validity.

Several points are important to consider in the design. The discriminant validity of the target ratings should be representative of the rating context so that accuracy findings generalize beyond the particular research setting. Brunswik's (1956) view of representative design underscores this point. Unfortunately, representative designs are apt to be expensive. Most accuracy studies have used videotapes of four to eight ratees who are each rated on six to 12 dimensions (e.g., Borman, 1977, 1979a; Hedge, 1982; McIntyre et al., 1984; Murphy et al., 1982). Such small combinations of ratees and dimensions restrict the amount of discriminant validity that can be designed into the target ratings and, therefore, restrict the generality of the research findings.

The combination design can be expanded to consider the broad

scope of research on performance ratings. Multiple raters can be included in the Rating Sources so as to include rater characteristics such as sex, race, ability, and motivation. Effects coding of the raters against the expert source will provide the statistics for each rater that are contained in the combination design (Kerlinger & Pedhazur, 1973). Ratee characteristics can also be studied in the combination design. Videotapes of actors can be constructed whose target ratings are identical but who differ in characteristics such as age, sex, and race. Furthermore, manipulations of ratee and rater characteristics in the same design address important legal questions about equal employment opportunity and the quality of performance ratings (Cascio & Bernardin, 1981). Finally, contextual factors can be evaluated for their impact on the accuracy of ratings. Factors that can be studied include the intended use of the ratings (McIntyre et al., 1984), the content of the attributes (Kavanagh, 1971), and the feedback given to raters on their accuracy (Ilgen, Fisher, & Taylor, 1979).

## VI. RECOMMENDATIONS

No research study has implemented the combination design to investigate both the validity and accuracy of performance ratings. The design provides a rich framework for understanding the distortions in performance ratings and can identify factors to control or remove to improve ratings. To date most research on the accuracy of ratings has focused on the training of raters to become more accurate in their ratings. Several programs have been used to train raters to make more accurate ratings. This line of research should continue; however, it must be expanded to address the influence of logical requirements for performance measures on accuracy training. For example, a study should be undertaken to consider the impact of accuracy training on performance measures that differ in their amounts of discriminant validity. The combination design developed in this paper provides an encompassing research strategy for future studies to evaluate the validity and accuracy of performance ratings.

REFERENCES


Bernardin, H. J., & Pence, E. C. (1980). The effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 258-272.

Borman, W. C. (1978). Exploring the upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 63, 135-144.

Borman, W. C. (1979a). Format and training effects on rating accuracy, and rater errors. Journal of Applied Psychology, 64, 410-421.

Borman, W. C. (1979b). Individual difference correlates of accuracy in evaluating others' performance effectiveness. Applied Psychological Measurement, 3, 103-115.

Borman, W. C., & Dunnette, M. D. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 60, 561-565.

Boruch, R. F., Larkin, J. D., Wolins, L., & MacKinney, A. C. (1970). Alternative methods of analysis: Multitrait-multimethod data. Educational and Psychological Measurement, 30, 833-853.

Brunswik, E. (1956). Perception and the representative design of psychological experiments (2nd ed.). Berkeley: University of Colorado Press.

Burnaska, R. F., & Hollmann, T. D. (1974). An empirical comparison of the relative effects of rater response biases on three rating scale formats. Journal of Applied Psychology, 59, 307-312.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cascio, W. F., & Bernardin, H. J. (1981). Implications of performance appraisal litigation for performance appraisal decisions. Personnel Psychology, 34, 211-226.

Cline, V. B., & Richards, J. M., Jr. (1960). Accuracy of interpersonal perception -- A general trait? Journal of Abnormal and Social Psychology, 60, 1-7.

Cronbach, L. J. (1955). Processes affecting scores on understanding of others and assuming "similarity." Psychological Bulletin, 52, 177-193.

Crow, W. J., & Hammond, K. R. (1957). The generality of accuracy and response sets in interpersonal perception. Journal of Abnormal and Social Psychology, 54, 384-390.

DeCotiis, T. A., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. Academy of Management Review, 3, 635-646.

Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. Journal of Applied Psychology, 65, 147-154.

Hedge, J. W. (1982). Improving the accuracy of performance evaluations: A comparison of three methods of performance appraiser training. Unpublished doctoral dissertation, Old Dominion University.

Ilgen, D. R., Fisher, C. D., & Taylor, S. M. (1979). Consequences of individual feedback on behavior in organizations. Journal of Applied Psychology, 64, 359-371.

Kavanagh, M. J. (1971). The content issue in performance appraisal: A review. Personnel Psychology, 24, 653-668.

Kavanagh, M. J., Borman, W. C., Hedge, J. W., & Gould, R. B. (1986). Job performance measurement classification scheme for validation research in the military (AFHRL-TP-85-51, AD-A164837). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 75, 34-49.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart and Winston.

Kirk, R. E. (1968). Experimental design: Procedures for the behavioral sciences. Belmont, CA: Wadsworth.

Landy, F., & Farr, J. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

Landy, F. J., & Trumbo, D. A. (1980). Psychology of work behavior. Homewood, IL: The Dorsey Press.

Lawler, E. E., III. (1967). The multi-trait-multi-rater approach to measuring managerial job performance. Journal of Applied Psychology, 51, 369-381.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67, 280-296.

McCormick, E. J. (1976). Job and task analysis. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67, 320-325.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.

Schwab, D. P., Heneman, H. G., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-562.

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Stanley, J. C. (1961). Analysis of unreplicated three-way classifications with applications to rater bias and trait independence. Psychometrika, 26, 205-219.

Taylor, J. B. (1968). Rating scales as measures of clinical judgment: A method for increasing scale reliability and sensitivity. Educational and Psychological Measurement, 28, 747-766.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of ratings. Personnel Psychology, 35, 521-551.

Winer, B. J. (1971). Statistical principles in experimental design. New York: McGraw-Hill.

Table 1. Summary Table for the Psychometric Interpretations of the Basic Multitrait-Multimethod Design

| Source | Psychometric interpretation |
|---|---|
| Traits (T) | Trait Bias |
| Methods (M) | Scale Bias |
| T x M | Trait by Scale Bias |
| Ratees (R) | Convergent Validity |
| R x T | Discriminant Validity |
| R x M | Method Bias |
| Error | Sampling and Measurement Errors |

Table 2. Example Data for Basic Multitrait-Multimethod Design

| | Methods | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | | 2 | | |
| | Traits | | | Traits | | |
| Test administrators | 1 | 2 | 3 | 1 | 2 | 3 |
| 1 | 4 | 7 | 2 | 3 | 6 | 3 |
| 2 | 3 | 5 | 1 | 3 | 5 | 4 |
| 3 | 7 | 9 | 6 | 6 | 8 | 6 |
| 4 | 6 | 6 | 2 | 4 | 5 | 3 |
| 5 | 5 | 5 | 1 | 4 | 4 | 4 |
| 6 | 8 | 2 | 5 | 5 | 5 | 7 |
| 7 | 4 | 1 | 1 | 3 | 4 | 5 |
| 8 | 6 | 3 | 4 | 6 | 2 | 2 |
| 9 | 7 | 5 | 2 | 8 | 6 | 4 |
| 10 | 7 | 1 | 1 | 4 | 2 | 2 |

Note. Trait 1 is maintaining procedures; Trait 2 is gaining rapport; and Trait 3 is presenting instructions. Method 1 is example-anchored, and Method 2 is checklist.

Table 3. Summary Table for the Analysis of the Data for the Basic Multitrait-Multimethod Design

| Source | df | MS | F-Ratio | VC | ICC |
|---|---|---|---|---|---|
| Traits (T) | 2 | 18.87 | 4.43* | .49 | .10 |
| Methods (M) | 1 | .82 | .55 | -.01 | .00 |
| T x M | 2 | 8.47 | 7.84** | .24 | .05 |
| Ratees (R) | 9 | 9.57 | 8.86** | 1.42 | .29 |
| R x T | 18 | 4.26 | 3.94** | 1.59 | .32 |
| R x M | 9 | 1.48 | 1.37 | .13 | .03 |
| Error | 18 | 1.08 | | 1.08 | |

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the source's coefficient was set to zero. VC, variance component; ICC, intraclass correlation coefficient.

*$p < .05$. **$p < .01$.

Table 4. Summary Table for Psychometric Interpretations of the One-Factor Design Beyond the Multitrait-Multimethod Design

| Source | Psychometric interpretation |
|---|---|
| Purposes (P) | Research Conditions |
| Ratees (R)/P | Convergent Validity Within Research Conditions |
| Traits (T) | Trait Bias |
| T x P | Trait Bias by Research Conditions |
| T x R/P | Discriminant Validity Within Research Conditions |
| Methods (M) | Scale Bias |
| M x P | Scale Bias by Purpose |
| M x R/P | Method Bias Within Research Conditions |
| T x M | Trait by Scale Bias |
| T x M x P | Trait by Scale Bias by Research Conditions |
| Error | Measurement and Sampling Errors |

Table 5. Summary Table for Analysis of Data for One-Factor Design Beyond the Multitrait-Multimethod Design

| Source | df | MS | F-Ratio | VC | ICC |
|---|---|---|---|---|---|
| Purposes (P) | 1 | 3.75 | .36 | -.11 | .00 |
| Ratees (R)/P | 8 | 10.30 | 11.32* | 1.56 | .34 |
| Traits (T) | 2 | 18.87 | 10.14* | .57 | .12 |
| T x P | 2 | 23.40 | 12.58* | .71 | .15 |
| T x R/P | 16 | 1.86 | 2.04 | .48 | .10 |
| Methods (M) | 1 | .82 | .55 | -.01 | .00 |
| M x P | 1 | 1.35 | .90 | .00 | .00 |
| M x R/P | 8 | 1.50 | 1.65 | .20 | .04 |
| T x M | 2 | 8.47 | 9.31* | .25 | .05 |
| T x M x P | 2 | 2.40 | 2.64 | .05 | .01 |
| Error | 16 | .91 | | .91 | |

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the source's coefficient was set to zero. VC, variance component; ICC, intraclass correlation coefficient.

*$p < .01$.

Table 6. Summary Table for the Psychometric Interpretations of the Basic Accuracy Design

| Source | Psychometric interpretation |
|---|---|
| Rating Sources (S) | Elevation Accuracy |
| Ratees (R) | Convergent Validity |
| Traits (T) | Trait Bias |
| S x R | Differential Elevation Accuracy (Differential Convergent Validity by Rating Sources) |
| S x T | Stereotype Accuracy |
| R x T | Discriminant Validity |
| S x R x T | Differential Accuracy (Differential Discriminant Validity by Rating Sources) |

Table 7. Summary Table of Psychometric Interpretations of the Combination Design

| Source | Psychometric interpretation |
|---|---|
| Rating Sources (S) | Elevation Accuracy |
| Ratees (R) | Convergent Validity |
| Traits (T) | Trait Bias |
| Methods (M) | Scale Bias |
| S x R | Differential Elevation Accuracy (Differential Convergent Validity by Rating Sources) |
| S x T | Stereotype Accuracy |
| S x M | Differential Scale Bias by Rating Sources |
| R x T | Discriminant Validity |
| R x M | Method Bias |
| T x M | Trait by Scale Bias |
| S x R x T | Differential Accuracy (Differential Discriminant Validity by Rating Sources) |
| S x R x M | Differential Elevation Accuracy by (Differential Method Bias by Rating Sources) |
| S x T x M | Differential Stereotype Accuracy by Methods |
| R x T x M | Differential Discriminant Validity by Methods |
| Error | Measurement and Sampling Errors |

Table 8. Summary Table for the Analysis of the Data for the Combination Design

| Source | df | MS | F-Ratio | VC | ICC |
|---|---|---|---|---|---|
| Rating Sources (S) | 1 | 3.75 | .40 | -.18 | .00 |
| Ratees (R) | 4 | 11.31 | 1.22 | .17 | .03 |
| Traits (T) | 2 | 18.87 | .80 | -.18 | .00 |
| Methods (M) | 1 | .82 | .49 | -.03 | .00 |
| S x R | 4 | 9.29 | 12.39* | 1.42 | .25 |
| S x T | 2 | 23.40 | 15.60* | 2.19 | .39 |
| S x M | 1 | 1.35 | 1.52 | .03 | .01 |
| R x T | 8 | 2.22 | 1.48 | .18 | .03 |
| R x M | 4 | 2.11 | 2.37 | .20 | .04 |
| T x M | 2 | 8.47 | 2.66 | .19 | .03 |
| S x R x T | 8 | 1.50 | 2.00 | .38 | .07 |
| S x R x M | 4 | .89 | 1.19 | .05 | .01 |
| S x T x M | 2 | 2.40 | 3.20 | .33 | .06 |
| R x T x M | 8 | 1.07 | 1.43 | .16 | .03 |
| Error | 8 | .75 | | .75 | |

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the source's coefficient was set to zero. VC, variance component; ICC, intraclass correlation coefficient.

*$p < .01$.

END
DTIC
9-86